Modelos Bayesianos

Taller de Procesamiento de Señales

Agenda

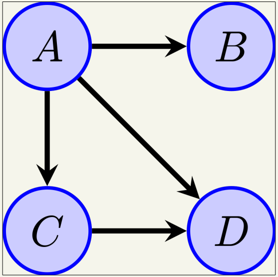
Redes Bayesianas

Modelos Gráficos

Modelos probabilísticos capaz de representarse con un grafo.

Red Bayesiana

Grafo acíclico dirigido que representa la relación de causalidad e independencia de sus variables. Dos variables aleatorias cualesquiera son condicionalmente independientes dados los valores de sus padres causales (y por lo tanto las raices son independientes).



$$p(A, B, C, D) = p(A) \cdot p(B|A) \cdot p(C|A) \cdot p(D|A, C)$$

Inferencia Bayesiana

- Los parámetros θ deben ser considerado realizaciones de una variable aleatoria T con una distribución a priori conocida $p(\theta)$.
- Las muestras son i.i.d. **cuando** se conoce el parámetro.
- La distribución a posteriori de los parámetros se calcula como:

$$p(\theta|\mathcal{D}_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

  con $\mathcal{D}_n = \{x_1, \cdots, x_n\}$.
- Como estimador puntual suele elegirse el *maximo a posteriori* (Θ discreto) y el estimador bayesiano o *media a posteriori* (Θ continuo).
- No son necesarios los estimadores puntuales para predecir:

$$p(x_{\text{test}}|\mathcal{D}_n) = \int_{\Theta} p(x_{\text{test}}|\theta) p(\theta|\mathcal{D}_n) d\theta = \mathbb{E}[p(x_{\text{test}}|T)|\mathcal{D}_n]$$

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

No confunduir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir "no sé"? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

Inferencia Bayesiana

Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

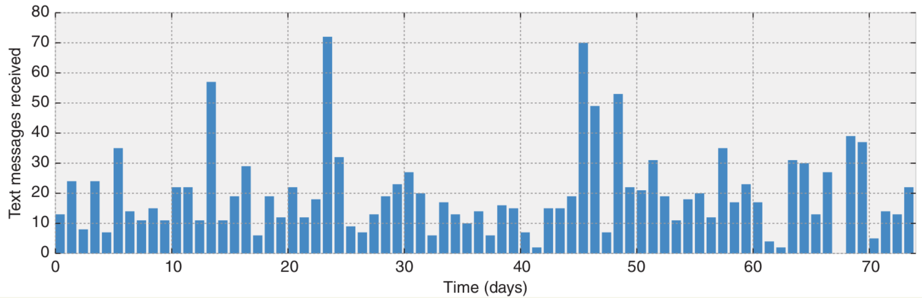
No confunduir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir "no sé"? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

¿Son prácticos los métodos Bayesianos?

Si, no solo por poder *adaptarse* a intentar resolver los mismos problemas que la estadística frecuentista (por ejemplo predicciones), sino que también pueden intentar resolver problemas donde la estadística clásica es insuficiente o iluminar el sistema subyacente con un modelado más flexible.

Ejemplo de modelado Bayesiano



Problema

Un usuario proporciona una serie de recuentos diarios de mensajes de whatsapp enviados. Tiene curiosidad por saber si los hábitos de envío de mensajes han cambiado con el tiempo. ¿Cómo puedes modelar esto?

Ejemplo de modelado Bayesiano

- La cantidad de mensajes en un día deberá ser modelada como una variable discreta cuyos átomos es $\mathbb{N}_0$. Por ejemplo $X_i \sim \text{Poi}(\lambda_i)$.
- Si observamos los datos, parecería que el valor de λ_i aumenta en algún momento durante las observaciones. ¿Cómo podemos representar matemáticamente esta observación? Supongamos que algún día τ durante el período de observación, el parámetro λ_i se incrementa repentinamente. Entonces realmente tenemos dos tasas: una para el período anterior a τ y otro para el resto del período:

$$\lambda_i = \begin{cases} \beta_1 & i < \tau \\ \beta_2 & i \geq \tau \end{cases}$$

- Tanto β_1 como β_2 toman valores reales no negativos. Por ejemplo $\beta_1, \beta_2 \sim \mathcal{E}(\alpha)$. Nuestra estimación de α no influye demasiado en el modelo, por lo que tenemos cierta flexibilidad en nuestra elección. Para evitar ser demasiado obstinados con este parámetro se sugiere:

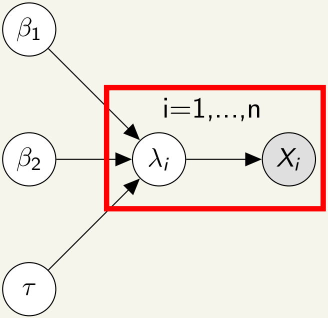
$$\alpha \approx \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Ejemplo de modelado Bayesiano

- ¿Qué pasa con τ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles $\tau \sim \mathcal{U}\{1 : T\}$.

Ejemplo de modelado Bayesiano

- ¿Qué pasa con τ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles $\tau \sim \mathcal{U}\{1 : T\}$.



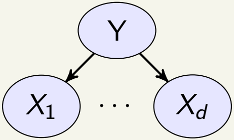
Outline

Naive Bayes

Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad $Y \to X$ con las diferentes componentes $X_j|Y = k$ independientes.

Red Bayesiana



Cálculo

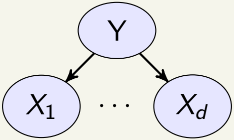
$$p(y|\mathrm{x}) = \frac{p(y)\prod_{j=1}^{d} p(x_j|y)}{p(\mathrm{x})}$$

Naive Bayes

Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad $Y \to X$ con las diferentes componentes $X_j|Y = k$ independientes.

Red Bayesiana



Cálculo

$$p(y|\mathrm{x}) = \frac{p(y)\prod_{j=1}^{d} p(x_j|y)}{p(\mathrm{x})}$$

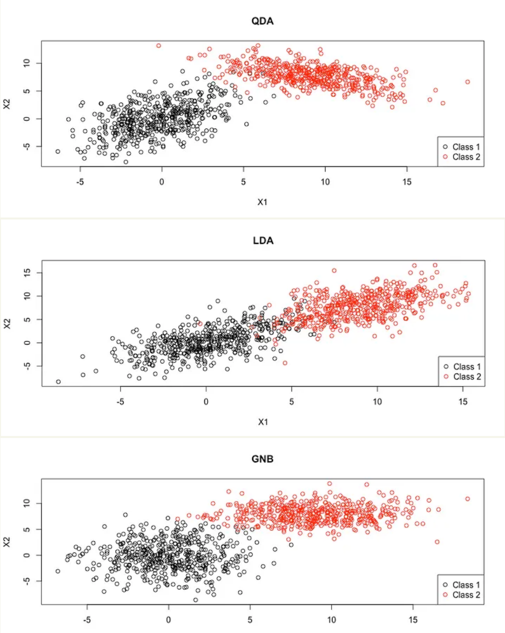
Gaussian Naive Bayes

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{(\mu_1^{(k)}, \sigma_1^{2(k)}), \cdots (\mu_d^{(k)}, \sigma_d^{2(k)})\}$, se modela $Y \sim \mathrm{Cat}(\{c_1, \cdots, c_K\})$ y $X_j|Y = k \sim \mathcal{N}(\mu_j^{(k)}, \sigma_j^{2(k)})$, para luego estimar los parámetros.

Gaussian Naive Bayes (GNB)

Diferencias entre QDA, LDA y GNB

- QDA acepta como Σ_k cualquer conjunto de matrices definidas positiva.
- LDA acepta como Σ_k cualquier matriz pero todas iguales.
- GNB permite tener matrices Σ_k diferentes pero todas diagonales.



Gaussian Naive Bayes (GNB)

QDA

$$\Sigma_k = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x - \mu^{(k)}) \left(x - \mu^{(k)}\right)^T$$

LDA

$$\Sigma = \frac{1}{n - K} \sum_{k=1}^{K} (|\mathcal{D}_k| - 1)\Sigma_k$$

GNB

$$\Sigma_k = \mathrm{DIAG}\left(\sigma_1^{2(k)}, \cdots, \sigma_d^{2(k)}\right), \qquad \sigma_j^{2(k)} = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x_j - \mu_j^{(k)})^2$$

Outline

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X_j | Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X_j | Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Inferencia

$$p(y|\mathrm{x}) \propto c_y \prod_{j=1}^{d} \theta_{x_j}^{(y)} = c_y \prod_{m=1}^{V} \left(\theta_m^{(y)}\right)^{N_m}$$

con N_m : Cantidad de predictores con valor m.

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB)

Sea $\pi = \{c_1, \cdots, c_K\}$ y $\theta^{(k)} = \{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\}$, se modela como $Y \sim \text{Cat}(\{c_1, \cdots, c_K\})$ y $X_j|Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \cdots, \theta_V^{(k)}\})$, utilizando estimadores puntuales.

Inferencia

$$p(y|\mathrm{x}) \propto c_y \prod_{j=1}^{d} \theta_{x_j}^{(y)} = c_y \prod_{m=1}^{V} \left(\theta_m^{(y)}\right)^{N_m}$$

con N_m : Cantidad de predictores con valor m.

Sobre las variables contadoras

Sea $\mathrm{N} = (N_1, \cdots, N_V)$, es sencillo notar que $\sum_{m=1}^{V} N_m = d$ y $\mathrm{N}|_{Y=k} \sim \mathcal{M}_n(d, [\theta_1^{(k)}, \cdots, \theta_V^{(k)}])$.

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_{y} \; p(y|\mathrm{x}) = \arg\max_{y} \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_y \; p(y|\mathrm{x}) = \arg\max_y \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Probabilidades de las Clases

Los parámetros $c_1, \cdots, c_K$ son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg\max_y \; p(y|\mathrm{x}) = \arg\max_y \; \log(c_y) + \sum_{m=1}^{V} N_m \log\left(\theta_m^{(y)}\right)$$

Probabilidades de las Clases

Los parámetros $c_1, \cdots, c_K$ son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

Sobre el valor d

Cada muestra puede poseer un valor de d diferente. Eso es típico en texto, donde cada documento posee una cantidad diferente de palabras.

Multinomial Naive Bayes

Estimación de $\theta_m^{(k)}$

Se cuenta con datos $\{(\mathrm{N}_i, y_i)\}_{i=1}^n$. Sin embargo, para cada clase k se utilizarán solamente los datos con $\{y_i = k\}$ distribuidos como una multinomial de probabilidades $\theta_1^{(k)}, \cdots, \theta_V^{(k)}$. A su vez, dado que las variables N_m cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^{n} N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

Multinomial Naive Bayes

Estimación de $\theta_m^{(k)}$

Se cuenta con datos $\{(\mathrm{N}_i, y_i)\}_{i=1}^n$. Sin embargo, para cada clase k se utilizarán solamente los datos con $\{y_i = k\}$ distribuidos como una multinomial de probabilidades $\theta_1^{(k)}, \cdots, \theta_V^{(k)}$. A su vez, dado que las variables N_m cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^{n} N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

Modelado: Estimador Bayesiano

Como modelado para el entrenamiento se supone $\mathrm{T} \sim \mathrm{Dir}([\alpha_1, \cdots, \alpha_V])$ y $(\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)})|_{\mathrm{T}=\Theta} \sim \mathcal{M}_n(\tilde{d}^{(k)}, [\theta_1^{(k)}, \cdots, \theta_V^{(k)}])$.

Multinomial Naive Bayes

Dirichlett Distribution

El vector aleatorio $(T_1, \cdots, T_V) \sim \text{Dir}([\alpha_1, \cdots, \alpha_V])$ puede ser pensado como una beta multivariada. Su densidad es de la forma

$$p(\theta_1, \cdots, \theta_V) = \frac{\prod_{m=1}^{V} \Gamma(\alpha_m)}{\Gamma\left(\sum_{m=1}^{V} \alpha_m\right)} \left(\prod_{m=1}^{V} \theta_m^{\alpha_m - 1}\right) \cdot \mathbb{1}\left\{\sum_{m=1}^{V} \theta_m = 1, \theta_m \geq 0\right\}$$

con sus marginales $T_m \sim \beta(\alpha_m, \sum_{\eta \neq m} \alpha_\eta)$.

Sobre la beta

Recordar que si $T \sim \beta(a, b)$, entonces $\mathbb{E}[T] = \frac{a}{a+b}$.

Multinomial Naive Bayes

Distribución a Posteriori

$$
\begin{aligned}
&p\left(\theta_1^{(k)},\cdots,\theta_V^{(k)}|\tilde{N}_1^{(k)},\cdots,\tilde{N}_V^{(k)}\right)\\
&\quad\propto P\left(\tilde{N}_1^{(k)},\cdots,\tilde{N}_V^{(k)}|\theta_1^{(k)},\cdots,\theta_V^{(k)}\right)\cdot p\left(\theta_1^{(k)},\cdots,\theta_V^{(k)}\right)\\
&\quad\propto\left(\prod_{m=1}^{V}(\theta_m^{(k)})^{\tilde{N}_m^{(k)}}\right)\left(\prod_{m=1}^{V}(\theta_m^{(k)})^{\alpha_m-1}\cdot\mathbb{1}\left\{\theta_m^{(k)}\geq 0\right\}\right)\cdot\mathbb{1}\left\{\sum_{m=1}^{V}\theta_m^{(k)}=1\right\}
\end{aligned}
$$

con lo cuál $\mathrm{T}|_{\tilde{N}_1^{(k)},\cdots,\tilde{N}_V^{(k)}}\sim\mathrm{Dir}([\tilde{N}_1^{(k)}+\alpha_1,\cdots,\tilde{N}_V^{(k)}+\alpha_V])$

Multinomial Naive Bayes

Distribución a Posteriori

$$
\begin{aligned}
& p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)} | \tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}\right) \\
& \quad \propto P\left(\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)} | \theta_1^{(k)}, \cdots, \theta_V^{(k)}\right) \cdot p\left(\theta_1^{(k)}, \cdots, \theta_V^{(k)}\right) \\
& \quad \propto \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\tilde{N}_m^{(k)}}\right) \left(\prod_{m=1}^{V} (\theta_m^{(k)})^{\alpha_m - 1} \cdot \mathbb{1}\left\{\theta_m^{(k)} \geq 0\right\}\right) \cdot \mathbb{1}\left\{\sum_{m=1}^{V} \theta_m^{(k)} = 1\right\}
\end{aligned}
$$

con lo cuál $\mathrm{T}|_{\tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}} \sim \mathrm{Dir}([\tilde{N}_1^{(k)} + \alpha_1, \cdots, \tilde{N}_V^{(k)} + \alpha_V])$

Estimador Bayesiano

$$
\hat{\theta}_m^{(k)} = \mathbb{E}[T_m | \tilde{N}_1^{(k)}, \cdots, \tilde{N}_V^{(k)}] = \frac{\tilde{N}_m^{(k)} + \alpha_m}{\sum_{\eta=1}^{V} \tilde{N}_\eta^{(k)} + \alpha_\eta}
$$

Outline

Variational Bayes

Variational Bayes

La idea es considerar los parámetros como parte del espacio latente. Sea Z un vector no observable del problema, en general será prohibitivo calcular la distribución a posteriori $p(\mathrm{z}|\mathrm{x})$. Con lo cuál, uno aproximará dicha distribución con la solución de

$$\begin{aligned}\arg\min_{q\in\mathcal{P}} \mathrm{KL}\left(q(\cdot|\mathrm{x})\|p(\cdot|\mathrm{x})\right) &= \arg\max_{q\in\mathcal{P}} H(q(\cdot|\mathrm{x})) + \mathbb{E}_q\left[\log p(\mathrm{x},\mathrm{Z})|\mathrm{X}=\mathrm{x}\right] \\ &= \arg\max_{q\in\mathcal{P}} \mathrm{ELBO}(q)\end{aligned}$$

donde $q(\mathrm{z}|\mathrm{x})$ cumple ciertas restricciones $\mathcal{P}$.

Variational Bayes

Variational Bayes

La idea es considerar los parámetros como parte del espacio latente. Sea Z un vector no observable del problema, en general será prohibitivo calcular la distribución a posteriori $p(\mathrm{z}|\mathrm{x})$. Con lo cuál, uno aproximará dicha distribución con la solución de

$$\begin{aligned}\arg\min_{q\in\mathcal{P}} \mathrm{KL}\left(q(\cdot|\mathrm{x})\|p(\cdot|\mathrm{x})\right) &= \arg\max_{q\in\mathcal{P}} H(q(\cdot|\mathrm{x})) + \mathbb{E}_q\left[\log p(\mathrm{x},\mathrm{Z})|\mathrm{X}=\mathrm{x}\right] \\ &= \arg\max_{q\in\mathcal{P}} \mathrm{ELBO}(q)\end{aligned}$$

donde $q(\mathrm{z}|\mathrm{x})$ cumple ciertas restricciones $\mathcal{P}$.

Mean field approximation

Aproximación que relaja el problema al suponer que q se puede factorizar como productos de densidades tratables. Por ejemplo, sea $\mathrm{z}=(\mathrm{u},\phi)$ se relaja el problema suponiendo $q(\mathrm{z}|\mathrm{x})=q_1(\mathrm{u}|\mathrm{x})q_2(\phi|\mathrm{x})$ para todo $q\in\mathcal{P}$.

Variational Bayes

Planteo del problema

Se buscan q_1 y q_2 que maximicen

$$
\begin{aligned}
&\mathrm{ELBO}(q) \\
&= H(q_1(\cdot|\mathrm{x})) + H(q_2(\cdot|\mathrm{x})) + \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \left(\int_{\Phi} q_2(\phi|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\phi \right) d\mathrm{u} \\
&= H(q_1(\cdot|\mathrm{x})) + H(q_2(\cdot|\mathrm{x})) + \int_{\Phi} q_2(\phi|\mathrm{x}) \left(\int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\mathrm{u} \right) d\phi
\end{aligned}
$$

Variational Bayes

Planteo del problema

Se buscan q_1 y q_2 que maximicen

$$
\begin{aligned}
&\text{ELBO}(q)\\
&= H(q_1(\cdot|\mathrm{x})) + H(q_2(\cdot|\mathrm{x})) + \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \left(\int_{\Phi} q_2(\phi|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\phi \right) d\mathrm{u}\\
&= H(q_1(\cdot|\mathrm{x})) + H(q_2(\cdot|\mathrm{x})) + \int_{\Phi} q_2(\phi|\mathrm{x}) \left(\int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\mathrm{u} \right) d\phi
\end{aligned}
$$

Resolución Iterativa

Se simplificará el problema resolviéndolo de forma iterativa. Definimos

$$
\begin{aligned}
E_1(\mathrm{x}, \mathrm{u}) &= \int_{\Phi} q_2(\phi|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\phi \equiv f(q_2)\\
E_2(\mathrm{x}, \phi) &= \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log p(\mathrm{x}, \mathrm{u}, \phi) d\mathrm{u} \equiv f(q_1)
\end{aligned}
$$

Variational Bayes

Problemas

$$q_1(\cdot|\mathrm{x}) = \arg\max_{q_1} \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_1(\mathrm{x},\mathrm{u})}}{q_1(\mathrm{u}|\mathrm{x})} d\mathrm{u}$$

$$q_2(\cdot|\mathrm{x}) = \arg\max_{q_2} \int_{\Phi} q_2(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_2(\mathrm{x},\phi)}}{q_2(\phi|\mathrm{x})} d\phi$$

Variational Bayes

Problemas

$$q_1(\cdot|\mathrm{x}) = \arg\max_{q_1} \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_1(\mathrm{x},\mathrm{u})}}{q_1(\mathrm{u}|\mathrm{x})} d\mathrm{u}$$
$$q_2(\cdot|\mathrm{x}) = \arg\max_{q_2} \int_{\Phi} q_2(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_2(\mathrm{x},\phi)}}{q_2(\phi|\mathrm{x})} d\phi$$

Soluciones

La solución consiste en iterar entre

$$q_1(\mathrm{u}|\mathrm{x}) \propto e^{E_1(\mathrm{x},\mathrm{u})}, \qquad q_2(\phi|\mathrm{x}) \propto e^{E_2(\mathrm{x},\phi)}$$

Variational Bayes

Problemas

$$q_1(\cdot|\mathrm{x}) = \arg\max_{q_1} \int_{\mathcal{U}} q_1(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_1(\mathrm{x},\mathrm{u})}}{q_1(\mathrm{u}|\mathrm{x})} d\mathrm{u}$$
$$q_2(\cdot|\mathrm{x}) = \arg\max_{q_2} \int_{\Phi} q_2(\mathrm{u}|\mathrm{x}) \log \frac{e^{E_2(\mathrm{x},\phi)}}{q_2(\phi|\mathrm{x})} d\phi$$

Soluciones

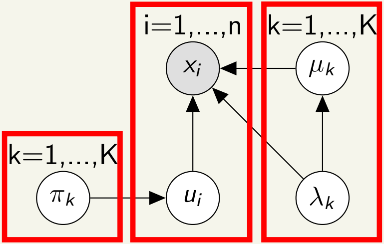
La solución consiste en iterar entre

$$q_1(\mathrm{u}|\mathrm{x}) \propto e^{E_1(\mathrm{x},\mathrm{u})}, \qquad q_2(\phi|\mathrm{x}) \propto e^{E_2(\mathrm{x},\phi)}$$

Sobre el ELBO

$$\mathrm{ELBO}(q) = \log p(\mathrm{x}) - \mathrm{KL}\left(q(\cdot|\mathrm{x}) \| p(\cdot|\mathrm{x})\right) \leq \log p(\mathrm{x})$$
$$\mathrm{ELBO}(q) = \mathbb{E}_q\left[\log p(\mathrm{x}|\mathrm{z})|\mathrm{x}\right] - \mathrm{KL}\left(q(\cdot|\mathrm{x}) \| p(\cdot)\right) \leq \mathbb{E}_q\left[\log p(\mathrm{x}|\mathrm{z})|\mathrm{x}\right]$$

Gaussian Variational Bayes



Modelo

$$p(\mathrm{x}, \mathrm{u}, \pi, \lambda, \mu) = p(\pi)\left(\prod_{k=1}^{K} p(\lambda_k) p(\mu_k|\lambda_k)\right)\left(\prod_{i=1}^{n} P(u_i|\pi) p(x_i|u_i, \mu, \lambda)\right)$$

con

$$\pi \sim \mathrm{Dir}(\alpha), \qquad \lambda_k \sim \Gamma(\nu, \beta), \qquad \mu_k|\lambda_k \sim \mathcal{N}(m, (\delta\lambda_k)^{-1})$$
$$u|\pi \sim \mathrm{Cat}(\pi), \qquad x|u, \mu, \lambda \sim \mathcal{N}(\mu_u, \lambda_u^{-1})$$

Gaussian Variational Bayes

Mean field approximation

Se aproxima $q(\mathrm{u}, \pi, \lambda, \mu|\mathrm{x}) = Q_1(\mathrm{u}|\mathrm{x})q_2(\pi, \lambda, \mu|\mathrm{x})$ y luego

$$E_1(\mathrm{x}, \mathrm{u}) = \mathrm{cte} + \sum_{i=1}^{n} \int q_2(\pi|\mathrm{x}) \log P(u_i|\pi) d\pi$$

$$+ \sum_{i=1}^{n} \int \int q_2(\mu, \lambda|\mathrm{x}) \log p(x_i|u_i, \mu, \lambda) d\mu d\lambda$$

$$E_2(\mathrm{x}, \pi, \lambda, \mu) = \log p(\pi) + \sum_{k=1}^{K} \log p(\lambda_k) + \sum_{k=1}^{K} \log p(\mu_k|\lambda_k)$$

$$+ \sum_{i=1}^{n} \sum_{k=1}^{K} Q_1(u_i = k|\mathrm{x}) \log P(u_i = k|\pi)$$

$$+ \sum_{i=1}^{n} \sum_{k=1}^{K} Q_1(u_i = k|\mathrm{x}) \log p(x_i|u_i = k, \mu, \lambda)$$

Gaussian Variational Bayes

Cómputo de $q_2(\pi, \lambda, \mu|\mathrm{x})$

Lo primero a notar es la factorización. Sea $\gamma_{i,k} = Q_1(u_i = k|\mathrm{x})$, luego

$$q_2(\pi, \lambda, \mu|\mathrm{x}) \propto p(\pi) \left(\prod_{k=1}^{K} p(\lambda_k) p(\mu_k|\lambda_k) \right) \prod_{k=1}^{K} e^{\sum_{i=1}^{n} \gamma_{i,k} [\log \pi_k + \log \mathcal{N}_{x_i}(\mu_k, \lambda_k^{-1})]}$$

Luego $q_2(\pi, \lambda, \mu|\mathrm{x}) = q_2(\pi|\mathrm{x}) \prod_{k=1}^{K} q_2(\mu_k, \lambda_k|\mathrm{x})$.

Gaussian Variational Bayes

Cómputo de $q_2(\pi, \lambda, \mu|\mathrm{x})$

Lo primero a notar es la factorización. Sea $\gamma_{i,k} = Q_1(u_i = k|\mathrm{x})$, luego

$$q_2(\pi, \lambda, \mu|\mathrm{x}) \propto p(\pi) \left(\prod_{k=1}^{K} p(\lambda_k) p(\mu_k|\lambda_k) \right) \prod_{k=1}^{K} e^{\sum_{i=1}^{n} \gamma_{i,k} [\log \pi_k + \log \mathcal{N}_{x_i}(\mu_k, \lambda_k^{-1})]}$$

Luego $q_2(\pi, \lambda, \mu|\mathrm{x}) = q_2(\pi|\mathrm{x}) \prod_{k=1}^{K} q_2(\mu_k, \lambda_k|\mathrm{x})$.

Cómputo de $q_2(\pi|\mathrm{x})$

$$q_2(\pi|\mathrm{x}) \propto \left(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} e^{\sum_{i=1}^{n} \gamma_{i,k} \log \pi_k} \right) \mathbb{1} \left\{ \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0 \right\}$$

con lo cual $\pi|\mathrm{x} \sim \mathrm{Dir}([\alpha_1 + \sum_{i=1}^{n} \gamma_{i,1}, \cdots, \alpha_K + \sum_{i=1}^{n} \gamma_{i,K}])$

Gaussian Variational Bayes

Cómputo de $q_2(\mu_k, \lambda_k|\mathrm{x})$

$$q_2(\mu_k, \lambda_k|\mathrm{x})$$

$$\propto \underbrace{\lambda_k^{\nu-1} e^{-\beta\lambda_k} \mathbb{1}\{\lambda_k > 0\}}_{\propto p(\lambda_k)} \underbrace{\lambda_k^{1/2} e^{\frac{-\delta\lambda_k(\mu_k - m)^2}{2}}}_{\propto p(\mu_k|\lambda_k)} \lambda_k^{\frac{1}{2}\sum_{i=1}^n \gamma_{i,k}} e^{\frac{-\lambda_k \sum_{i=1}^n \gamma_{i,k}(x_i - \mu_k)^2}{2}}$$

con lo cual $\mu_k|\lambda_k, \mathrm{x} \sim \mathcal{N}\left(\frac{\delta m + \sum_{i=1}^n \gamma_{i,k} x_i}{\delta + \sum_{i=1}^n \gamma_{i,k}}, \frac{1}{\lambda_k\left(\delta + \sum_{i=1}^n \gamma_{i,k}\right)}\right)$ y
$\lambda_k|\mathrm{x} \sim \Gamma\left(\nu + \frac{1}{2}\sum_{i=1}^n \gamma_{i,k}, \beta + \frac{\delta m^2}{2} + \frac{1}{2}\sum_{i=1}^n \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^n \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^n \gamma_{i,k})}\right)$

Gaussian Variational Bayes

Cómputo de $q_2(\mu_k, \lambda_k|\mathrm{x})$

$$q_2(\mu_k, \lambda_k|\mathrm{x})$$

$$\propto \underbrace{\lambda_k^{\nu-1} e^{-\beta\lambda_k} \mathbb{1}\{\lambda_k > 0\}}_{\propto p(\lambda_k)} \underbrace{\lambda_k^{1/2} e^{\frac{-\delta\lambda_k(\mu_k - m)^2}{2}}}_{\propto p(\mu_k|\lambda_k)} \lambda_k^{\frac{1}{2}\sum_{i=1}^n \gamma_{i,k}} e^{\frac{-\lambda_k \sum_{i=1}^n \gamma_{i,k}(x_i - \mu_k)^2}{2}}$$

con lo cual $\mu_k|\lambda_k, \mathrm{x} \sim \mathcal{N}\left(\frac{\delta m + \sum_{i=1}^n \gamma_{i,k} x_i}{\delta + \sum_{i=1}^n \gamma_{i,k}}, \frac{1}{\lambda_k\left(\delta + \sum_{i=1}^n \gamma_{i,k}\right)}\right)$ y
$\lambda_k|\mathrm{x} \sim \mathit{\Gamma}\left(\nu + \frac{1}{2}\sum_{i=1}^n \gamma_{i,k}, \beta + \frac{\delta m^2}{2} + \frac{1}{2}\sum_{i=1}^n \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^n \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^n \gamma_{i,k})}\right)$

Estadísticos Suficientes

Basta con computar $N_k = \sum_{i=1}^{n} \gamma_{i,k}$, $f_k = \sum_{i=1}^{n} \gamma_{i,k} x_i$ y $s_k = \sum_{i=1}^{n} \gamma_{i,k} x_i^2$.

Gaussian Variational Bayes

Propiedad de las distribuciones Gamma y Dirichlett

Sean $\lambda \sim \Gamma(\nu, \beta)$ y $\pi \sim \mathrm{Dir}(\alpha)$, se puede demostrar que

- $\mathbb{E}[\log \lambda] = \psi(\nu) - \log(\beta)$
- $\mathbb{E}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{c=1}^{K} \alpha_c\right)$

donde $\psi(\cdot)$ es la función digamma $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$.

Gaussian Variational Bayes

Propiedad de las distribuciones Gamma y Dirichlett

Sean $\lambda \sim \Gamma(\nu, \beta)$ y $\pi \sim \text{Dir}(\alpha)$, se puede demostrar que

- $\mathbb{E}[\log \lambda] = \psi(\nu) - \log(\beta)$
- $\mathbb{E}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{c=1}^{K} \alpha_c\right)$

donde $\psi(\cdot)$ es la función digamma $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$.

Cómputo de $Q_1(\mathrm{u}|\mathrm{x})$

Lo primero a notar es la independencia

$$
\begin{aligned}
Q_1(\mathrm{u}|\mathrm{x}) &\propto \prod_{i=1}^{n} e^{\int q_2(\pi|\mathrm{x}) \log P(u_i|\pi) d\pi + \int \int q_2(\mu,\lambda|\mathrm{x}) \log p(x_i|u_i,\mu,\lambda) d\mu d\lambda} \\
&= \prod_{i=1}^{n} Q_1(u_i|\mathrm{x})
\end{aligned}
$$

Gaussian Variational Bayes

Cómputo de $Q_1(u_i = k|\mathrm{x})$

Sean los parámetros de q_2 definidos como $\pi|\mathrm{x} \sim \mathrm{Dir}(\alpha^*)$, $\mu_k|\lambda_k, \mathrm{x} \sim \mathcal{N}\left(m_k^*, (\delta_k^* \lambda_k)^{-1}\right)$ y $\lambda_k|\mathrm{x} \sim \Gamma\left(\nu_k^*, \beta_k^*\right)$. Luego

$$Q_1(u_i = k|\mathrm{x}) \propto e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^K \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2}\mathbb{E}_{q_2}[\lambda_k(x_i - \mu_k)^2]}$$

con

$$\begin{aligned}
\mathbb{E}_{q_2}[\lambda_k(x_i - \mu_k)^2] &= \mathbb{E}_{q_2}[\lambda_k \mathbb{E}_{q_2}[(x_i - \mu_k)^2|\lambda_k]] \\
&= \mathbb{E}_{q_2}[\lambda_k(\mathbb{E}_{q_2}[(\mu_k - m_k^*)^2|\lambda_k] + (m_k^* - x_i)^2)] \\
&= \frac{1}{\delta_k^*} + \frac{\nu_k^*}{\beta_k^*}(m_k^* - x_i)^2
\end{aligned}$$

Finalmente

$$Q_1(u_i = k|\mathrm{x}) \propto e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^K \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2\delta_k^*} - \frac{\nu_k^*}{2\beta_k^*}(m_k^* - x_i)^2}$$

Gaussian Variational Bayes

Solución: Inicializar $\gamma_{i,k}$ con EM e iterar entre

- Calcular $(\alpha_k^*, m_k^*, \delta_k^*, \nu_k^*, \beta_k^*)$ a partir de $\gamma_{i,k}$ como

$$\alpha_k^* = \alpha_k + \sum_{i=1}^{n} \gamma_{i,k}, \quad m_k^* = \frac{\delta m + \sum_{i=1}^{n} \gamma_{i,k} x_i}{\delta + \sum_{i=1}^{n} \gamma_{i,k}}$$

$$\delta_k^* = \delta + \sum_{i=1}^{n} \gamma_{i,k}, \quad \nu_k^* = \nu + \frac{1}{2} \sum_{i=1}^{n} \gamma_{i,k}$$

$$\beta_k^* = \beta + \frac{\delta m^2}{2} + \frac{1}{2} \sum_{i=1}^{n} \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^{n} \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^{n} \gamma_{i,k})}$$

- Calcular $\gamma_{i,k} = \frac{\rho_{i,k}}{\sum_{c=1}^{K} \rho_{i,c}}$ a partir de $(\alpha^*, m_k^*, \delta_k^*, \nu_k^*, \beta_k^*)$ como

$$\rho_{i,k} = e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^{K} \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2\delta_k^*} - \frac{\nu_k^*}{2\beta_k^*}(m_k^* - x_i)^2}$$

Gaussian Variational Bayes

Predictiva: Es una mezcla!

$$
\begin{aligned}
p(x_{\text{test}}|\mathcal{D}_n) &= \mathbb{E}[p(x_{\text{test}}|\phi)|\mathcal{D}_n] \\
&= \sum_{k=1}^{K} \mathbb{E}[\pi_k|\mathcal{D}_n] \cdot \mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2} \middle| \mathcal{D}_n\right] \\
&= \sum_{k=1}^{K} \frac{\alpha_k^*}{\sum_{c=1}^{K} \alpha_c^*} \cdot \tilde{p}_k(x_{\text{test}}|\mathcal{D}_n)
\end{aligned}
$$

Gaussian Variational Bayes

Predictiva: Es una mezcla!

$$\begin{aligned} p(x_{\text{test}}|\mathcal{D}_n) &= \mathbb{E}[p(x_{\text{test}}|\phi)|\mathcal{D}_n] \\ &= \sum_{k=1}^{K} \mathbb{E}[\pi_k|\mathcal{D}_n] \cdot \mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2} \middle| \mathcal{D}_n\right] \\ &= \sum_{k=1}^{K} \frac{\alpha_k^*}{\sum_{c=1}^{K} \alpha_c^*} \cdot \tilde{p}_k(x_{\text{test}}|\mathcal{D}_n) \end{aligned}$$

Normal-Gamma Distribution

$$1 = \int_0^\infty \int_{-\infty}^\infty \frac{\beta^\nu}{\Gamma(\nu)} \lambda^{\nu-1} e^{-\beta\lambda} \sqrt{\frac{\delta\lambda}{2\pi}} e^{\frac{-\delta\lambda}{2}(\mu-m)^2} d\mu d\lambda$$

$$\int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \lambda^{\nu-\frac{1}{2}} e^{\frac{-\delta\lambda}{2}(\mu-m)^2-\lambda\beta} d\mu d\lambda = \frac{\Gamma(\nu)}{\beta^\nu\sqrt{\delta}}$$

Gaussian Variational Bayes

Calculo auxiliar de las densidades a mezclar

$$\mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2} \middle| \mathcal{D}_n\right]$$

$$= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\lambda}{2\pi}} e^{\frac{-\lambda}{2}(x_{\text{test}}-\mu)^2} \frac{\beta_k^{*\,\nu_k^*}\sqrt{\delta_k^*}}{\sqrt{2\pi}\Gamma(\nu_k^*)} \lambda^{\nu_k^*-\frac{1}{2}} e^{\frac{-\delta_k^*\lambda}{2}(\mu-m_k^*)^2-\lambda\beta_k^*} d\mu d\lambda$$

$$\propto \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{1}{2\pi}} \lambda^{\nu_k^*} e^{\frac{-\lambda}{2}(x_{\text{test}}-\mu)^2-\frac{\delta_k^*\lambda}{2}(\mu-m_k^*)^2-\lambda\beta_k^*} d\mu d\lambda$$

Gaussian Variational Bayes

Calculo auxiliar de las densidades a mezclar

$$\mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2} \middle| \mathcal{D}_n\right]$$

$$= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\lambda}{2\pi}} e^{\frac{-\lambda}{2}(x_{\text{test}}-\mu)^2} \frac{\beta_k^{*\nu_k^*}\sqrt{\delta_k^*}}{\sqrt{2\pi}\Gamma(\nu_k^*)} \lambda^{\nu_k^*-\frac{1}{2}} e^{\frac{-\delta_k^*\lambda}{2}(\mu-m_k^*)^2-\lambda\beta_k^*} d\mu d\lambda$$

$$\propto \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{1}{2\pi}} \lambda^{\nu_k^*} e^{\frac{-\lambda}{2}(x_{\text{test}}-\mu)^2-\frac{\delta_k^*\lambda}{2}(\mu-m_k^*)^2-\lambda\beta_k^*} d\mu d\lambda$$

Sustitución dentro de la integral

$$\tilde{\nu} = \nu_k^* + \frac{1}{2}, \qquad \tilde{\delta} = \delta_k^* + 1, \qquad \tilde{m} = \frac{x_{\text{test}} + \delta_k^* m_k^*}{\delta_k^* + 1}$$

$$\tilde{\beta} = \beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)}$$

Gaussian Variational Bayes

Calculo auxiliar de las densidades a mezclar

$$\mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}}e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2}\middle|\mathcal{D}_n\right] \propto \left(\beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^*+1)}\right)^{-(\nu_k^*+1/2)}$$

$$\propto \left(1 + \frac{\delta_k^*\nu_k^*}{(\delta_k^*+1)\beta_k^*}\frac{(x_{\text{test}} - m_k^*)^2}{2\nu_k^*}\right)^{-\frac{2\nu_k^*+1}{2}}$$

Gaussian Variational Bayes

Calculo auxiliar de las densidades a mezclar

$$\mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}}e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2}\middle|\mathcal{D}_n\right] \propto \left(\beta_k^* + \frac{\delta_k^*(x_{\text{test}}-m_k^*)^2}{2(\delta_k^*+1)}\right)^{-(\nu_k^*+1/2)}$$

$$\propto \left(1 + \frac{\delta_k^*\nu_k^*}{(\delta_k^*+1)\beta_k^*}\frac{(x_{\text{test}}-m_k^*)^2}{2\nu_k^*}\right)^{-\frac{2\nu_k^*+1}{2}}$$

Distribución t-Student Generalizada: $X \sim t(\mu, \Lambda, \nu)$ si

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\sqrt{\frac{\Lambda}{\pi\nu}}\left(1 + \Lambda\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Gaussian Variational Bayes

Calculo auxiliar de las densidades a mezclar

$$\mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}}e^{\frac{-\lambda_k}{2}(x_{\text{test}}-\mu_k)^2}\middle|\mathcal{D}_n\right] \propto \left(\beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)}\right)^{-(\nu_k^*+1/2)}$$

$$\propto \left(1 + \frac{\delta_k^*\nu_k^*}{(\delta_k^*+1)\beta_k^*}\frac{(x_{\text{test}} - m_k^*)^2}{2\nu_k^*}\right)^{-\frac{2\nu_k^*+1}{2}}$$

Distribución t-Student Generalizada: $X \sim t(\mu, \Lambda, \nu)$ si

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\sqrt{\frac{\Lambda}{\pi\nu}}\left(1 + \Lambda\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Predictiva

$$p(x_{\text{test}}|\mathcal{D}_n) = \sum_{k=1}^{K}\frac{\alpha_k^*}{\sum_{c=1}^{K}\alpha_c^*}\cdot t\left(m_k^*, \frac{\delta_k^*\nu_k^*}{(\delta_k^*+1)\beta_k^*}, 2\nu_k^*\right)$$

Outline

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Método de Monte-Carlo

Sean X_i variables aleatorias i.i.d. con pdf $p(x)$ o pmf $P(x)$, entonces

$$\int_{\mathbb{R}} g(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i), \qquad \sum_{x\in\mathbb{A}} g(x)P(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Monte-Carlo

Ley de los grandes números

Sean X_i variables aleatorias i.i.d. con esperanza finita $\mathbb{E}[X]$, entonces $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X]$ (w.p.1).

Método de Monte-Carlo

Sean X_i variables aleatorias i.i.d. con pdf $p(x)$ o pmf $P(x)$, entonces

$$\int_{\mathbb{R}} g(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i), \qquad \sum_{x\in\mathbb{A}} g(x)P(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Algunas variantes interesantes

- $\int_a^b g(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} g(X_i)$ con $X_i \sim \mathcal{U}(a,b)$.
- $\int_a^b g(x)dx \approx \frac{1}{kn}\sum_{i=1}^{n} \mathbb{1}\{a < X_i < b\}$ con X_i de pdf $p(x) = k\cdot g(x)$.
- $\int_{\mathbb{R}} g(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} g(X_i)$ con X_i de pdf $q(x)$.

Técnicas de Muestreo

Ideas Principales

- En lugar de obtener la distribución a posteiori, vamos a generar muestras de esta distribución.
- Vamos aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}}|\mathcal{D}_n) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x_{\text{test}}|T_i)$$

Técnicas de Muestreo

Ideas Principales

- En lugar de obtener la distribución a posteiori, vamos a generar muestras de esta distribución.
- Vamos aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}}|\mathcal{D}_n) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x_{\text{test}}|T_i)$$

Teorema de Ergodicidad

El problema principal es que no es sencillo generar muestras independientes de la posterior. El Teorema de Ergodicidad nos permite aplicar la ley de los grandes números, sin la hipótesis de independencia, para sucesiones de variables aleatorias con ciertas características.

$$\mathbb{E}[g(X)] \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

Técnicas de Muestreo

Proceso de Markov

$$T_0 \longrightarrow T_1 \longrightarrow T_2 \longrightarrow T_3 \longrightarrow T_4 \longrightarrow T_5 \longrightarrow \cdots$$

Una **cadena de Markov** es una sucesión de variables aleatorias $\{T_t\}_{t\in\mathbb{N}_0}$ que cumple la propiedad de *falta de memoria*: la distribución del próximo estado depende únicamente del estado actual, y no del pasado completo:

$$P(T_t = \theta \to T_{t+1} = \theta') = \pi(\theta'|\theta)$$

Técnicas de Muestreo

Ideas Principales

- En lugar de obtener la distribución a posteiori, vamos a generar muestras de esta distribución.
- Vamos aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}}|\mathcal{D}_n) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x_{\text{test}}|T_i)$$

Técnicas de Muestreo

Ideas Principales

- En lugar de obtener la distribución a posteiori, vamos a generar muestras de esta distribución.
- Vamos aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}}|\mathcal{D}_n) \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(x_{\text{test}}|T_i)$$

Muestreo de Gibbs

Supongamos que, debido a su complejidad, no podemos simular muestras de $p(x, y)$; pero que si es posible generar muestras de las condicionales $p(x|y)$ y $p(y|x)$. El muestreo de Gibbs consiste en, a partir de un x_0, iterar entre:

$$y_k \sim p(y|x_k), \qquad x_{k+1} \sim p(x|y_k)$$

Luego de suficientes simulaciones (resultado asintótico), el último par (x, y) estará distribuido (aproximadamente) por $p(x, y)$.

Técnicas de Muestreo

Markov Chain Monte-Carlo (MCMC)

$$x_0 \rightarrow y_0 \rightarrow x_1 \rightarrow y_1 \rightarrow x_2 \rightarrow y_2$$

Técnicas de Muestreo

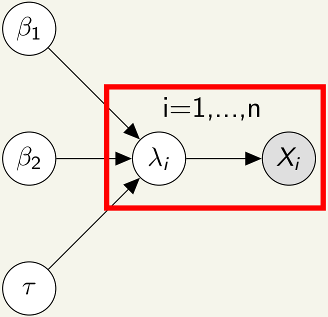
Markov Chain Monte-Carlo (MCMC)

$$x_0 \rightarrow y_0 \rightarrow x_1 \rightarrow y_1 \rightarrow x_2 \rightarrow y_2$$

Implementaciones más sofisticadas

1. Comenzar en la posición actual.
2. Proponer mudarse a una nueva posición cercana a la actual.
3. Aceptar/Rechazar la nueva posición basándose en el coherencia de la posición con los datos y distribuciones anteriores.
4.
    - Si acepta: Pasar a la nueva posición. Regresar al Paso 1.
    - De lo contrario: no moverse de la posición actual. Regrese al Paso 1.
5. Después de una gran cantidad de iteraciones, se reportan todas las posiciones aceptadas.

PYMC: Programación Probabilística con Monte-Carlo



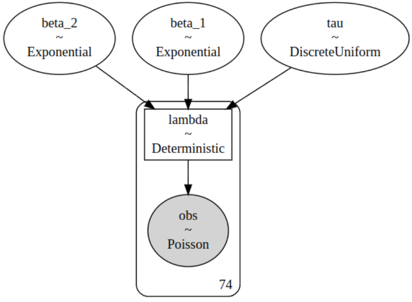
PYMC

```python
import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)


with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda",pm.math.switch(tau > idx, beta_1, beta_2))
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000,chains=2)

pm.model_to_graphviz(model)
```



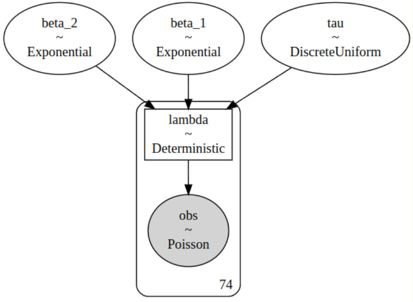
PYMC

```python
import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)


with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda",pm.math.switch(tau > idx, beta_1, beta_2))
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000,chains=2)

pm.model_to_graphviz(model)
```

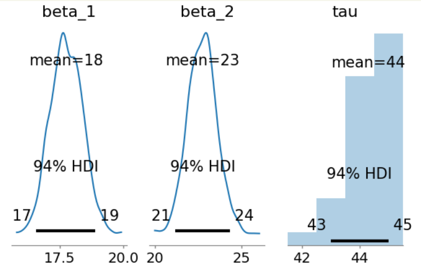



```python
beta_1_samples = trace.posterior['beta_1']
beta_2_samples = trace.posterior['beta_2']
tau_samples = trace.posterior['tau']
lambda_samples = trace.posterior['lambda']

with model:
  posterior_pred = pm.sample_posterior_predictive(trace)

pred_samples = posterior_pred.posterior_predictive['obs']

pm.plot_posterior(trace.posterior[['beta_1','beta_2','tau']],figsize=(7,4))
```



Anexo: Distribuciones

Normal(μ, σ^2)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gamma(ν, β)

$$p(x) = \frac{\beta^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\beta x} \mathbb{1}\{x > 0\}$$

T-Student generalizada(μ, Λ, ν)

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \sqrt{\frac{\Lambda}{\pi\nu}} \left(1 + \Lambda\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Dirichlett$(\alpha_1 \cdots, \alpha_K)$

$$p(x_1, \cdots, x_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \left(\prod_{k=1}^K x_k^{\alpha_k - 1}\right) \cdot \mathbb{1}\left\{\sum_{k=1}^K x_k = 1, x_k \geq 0\right\}$$